

DOCUMENT RESUME

ED 398 358

CE 070 962

TITLE Evaluation in HRD.
PUB DATE Mar 95
NOTE 4lp.; In: Academy of Human Resource Development (AHRD) Conference Proceedings (St. Louis, MO, March 2-5, 1995); see CE 070 958. For other conference symposia, see CE 070 959-981.
PUB TYPE Speeches/Conference Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adult Education; Continuing Education; *Evaluation Methods; *Feedback; Human Resources; *Labor Force Development; *Management Development; Models; Program Effectiveness; Program Evaluation; Research Needs; *Total Quality Management

ABSTRACT

These four papers are from a symposium that was facilitated by Nancy Dixon at the 1995 conference of the Academy of Human Resource Development (HRD). "Reliability Testing of Multirater Feedback" (Wim J. Nijhof, Anne Jager) focuses on testing the reliability and validity of the multirater feedback system. Findings are reported that conclude that trainers are better raters and formative and summative processes and assessments should be separated. "In Search of an Integrative Model for HRD Evaluation" (Elwood F. Holton, III) argues the four-level Kirkpatrick system of training evaluation is really a taxonomy of outcomes and that research is needed to develop an integrated measurement model. An initial model that accounts for the impact of the primary intervening variables is proposed. Key studies supporting the model are reviewed, and a research agenda is proposed. "Building Training on the Pillars: Applying Total Quality in HRD" (Douglas H. Smith, Jo D. Gallagher) begins with an overview of the five leading total quality management (TQM) writers. Four basic principles or pillars are derived and serve as the basis of the application of TQM within the instructional setting. Examples of the relationship of the four pillars to instructional delivery are presented. "Using 360-Degree Feedback to Evaluate Management Development: New Data, New Insights" (Gary N. McLean, Marcia R. Systma, Kathryn Kerwin-Ryberg) discusses the use of a 360-degree feedback instrument designed to facilitate management development to evaluate the effectiveness of a training event. (YLB)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to
improve reproduction quality.

• Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

W. J. Nijhof, E. Holton
D. Smith
G. McLean

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

ACADEMY OF HUMAN RESOURCE DEVELOPMENT

1995 Conference Proceedings

**March 2-5, 1995
St. Louis, MO**

[4] **EVALUATION IN HRD** Facilitator: Nancy Dixon, George Washington University

Reliability Testing of Multirater Feedback

Wim J. Nijhof, University of Twente (The Netherlands)

Anne Jager, University of Twente (The Netherlands)

In Search of an Integrative Model for HRD Evaluation

Elwood F. Holton, III, Louisiana State University

Building on Training Pillars: Applying TQM in HRD

Douglas H. Smith, Florida International University

Jo D. Gallagher, Florida International University

360 Degree Feedback Evaluation: New Data, New Insights

Gary N. McLean, University of Minnesota

BEST COPY AVAILABLE

Reliability Testing of Multirater Feedback

Wim J. Nijhof & Anne Jager
University of Twente, The Netherlands

Multirater feedback is an essential part of a seminar concept used by ARAMARK. Before testing the effectiveness of the seminars in terms of learning effects and transfer of training, the reliability and validity of multirating has to be tested. The conclusion is that trainers are better raters, Formative and summative processes and assessments should be separated.

ARAMARK is the top catering service company in the world. The Human Resource Management and HRDdivision of ARAMARK in Germany has developed a system of training and personal development based on a seminar concept in which multirating plays a vital role. Three types of seminars - ARAMARK Manager Rallyes I, II and III - have been developed and are carried out by a preferred supplier. This seminar concept started in an experimental form in 1993. ARAMARK wants to know how effective the seminar concept is and whether transfer of training takes place at different levels in the organization. Before measuring the effectiveness of a seminar concept, we need to know whether the multirater feedback system is reliable and valid. In this study of the ARA Manager Rallye during 1994, the scores of five ARA Manager seminars are analysed from a psychometric perspective. The quality of measurements is evaluated with a focus on the improvement of a measuring instrument. According to measurement theory, any measurement instrument should meet the general standard of (1) validity and (2) reliability. A test has validity if it measures what it purports to measure. The reliability of a test refers to the consistency of the evaluation results. In the study it refers to the consistency between different raters, or interrater reliability. To be valid, test scores should also be reliable.

The ARA Manager Rallye

The ARA Manager Rallye is an instrument used during ARA Manager seminars

- (1) to assess the individual and team performances of participants, and simultaneously
- (2) to train the assessment skills of participants.

By that, the ARA Manager Rallye tries to fulfill the four functions of

- (a) selection of participants who will not be able to finish the seminar successfully.
- (b) information of participants and trainers about achievement of seminar goals
- (c) motivation of participants to goal-orientation and to apply the ARA Systemtasks.
- (d) coaching of participants in the planning of other learning activities and in transfer of seminar content to the work setting.

One of the key components of the Manager Rallye is multirater feedback. This feedback is used for assessing individual performance, group performance, and for managers to learn and use assessment skills. This type of assessment is also used for personnel decisions, for outplacement, and succession planning. According to Bracken (1994) multirater systems can fail or succeed dependent on process design and planning, instrument development, instrument design, administration, feedback processing and reporting, and action planning as a result of feedback. Russel (1994): "The validity of these reasons is now considered suspect. Other trainees are not always accurate in their observations and this can cause faulty or misleading feedback". Russel indicates other causes for invalid feedback: trainees are not skilled at giving feedback, being more evaluative than objective, etc (Russell, 1994, p. 49). Audio and video techniques are better.

It is clear from the introduction that the focus will be on testing the feedback processing and reporting in terms of validity and reliability measurement. A fair and valid

assessment instrument must be reliable. Concerning reliability, in the ARA Manager Rallye differences between raters are thought to be a major possible source of error, since the end score in the majority of stages is determined by a lot of different raters. For example, in the different stages D, B, and C, the average of individual participant ratings (including self-ratings) determines 50 percent of the performance score. Ratings of the trainer, together with ratings of training professionals of the personnel departments and HRD in the stage D (*ARA Manager Duties*), determine(s) also 50 percent of the end score in these stages. In the study, the trainer together with training professionals of HRD and Personnel is called the trainer team. To be meaningful, the end score in each of the stages in which different raters give their individual judgments should be an average of ratings which are reasonably consistent between these different raters. That means both between participants, participants and trainer (team), and members of the trainer team.

Research Questions

Four research questions are formulated regarding the reliability and validity of the ARAMARK Manager Rallye:

- (1) Are the scores in the ARAMARK Manager Rallye consistent between
 - (a) ratings of participants and trainer team?
 - (b) ratings of participants?
 - (c) ratings of trainer and training professionals?
 - (d) self-ratings and ratings of members of the trainer team?
- (2) Are the scores in the ARA Manager Rallye consistent
 - (a) stages and end score of the seminar?
 - (b) different stages?
- (3) Do the scores in the ARA Manager Rallye achieve a sufficient amount of spread?
- (4) How could the quality of scores in the ARA Manager Rallye be improved?

The ARAMARK Manager Rallyes consist of different stages, mostly 6. Starting with a Check In (to assess prior knowledge) and ending with a Check Out (to test the skills and knowledge needed for a (middle) management position, the other stages cover different aspects of performance. To determine the interrater reliability, scores of one of these stages (*ARA Manager Aufgabe: stage D*) are analysed. This stage is common for all the seminars. Analyses are carried out at three different levels: (1) average stage scores, (2) stage scores, and (3) criterium scores. On each of the levels, Spearman rank correlation coefficients r_{sp} are determined by using SPSS/PC 4.0 to indicate the degree of association between raters.

The following classification regarding the interrater correlation coefficient is used:

$r_{sp} \geq .80$	: excellent
$.70 \leq r_{sp} < .80$	: good
$.60 \leq r_{sp} < .70$	: satisfactory
$r_{sp} < .60$	: unsatisfactory

Sample of cases

In the study the scores of two seminars ARA Manager 1, two seminars ARA Manager 2, and one seminar ARA Manager 3 are analysed. Their characteristics are shown in Table 1. Beforehand, it was clear that it would not be possible to compute a statistic indicating the interrater reliability of all seminars together. Since in each seminar different raters are present, and (somewhat) different rating scales are used, a decision was made to analyse each seminar separately. This would then concern the stages, in which different raters give their individual judgments. However, between these three different kinds of stages in each seminar also a lot of differences exist regarding the composition of subgroups, and the use of different rating scales in different stages. That's why it seemed more meaningful to analyse ratings from each stage separately. The only stage of these three which has the same structure in every seminar is stage D (*ARA Manager Aufgabe(Duties)*). Aside from this, in the other stages not enough cases are rated to compute a meaningful correlation coefficient. The ARA Manager Aufgabe is also the only stage in which both internal and external assessments take place. Internal assessments involve those actually participating

in the learning/teaching process having control over the assessment' (Bell and Harris, 1990). In the ARA Manager Rallye, the trainer and participants are the internal assessors. External assessments involve an outsider to judge performances. The training professionals of ARAMARK are external assessors. This seems to show enough evidence that the stage ARA Manager Aufgabe of each seminar is the stage most appropriate to answer research question 1.

Table 1: Characteristics of the seminars

Seminar	Period	Trainer	N of participants ^a	N of training professionals	N of stages
ARA Manager 1	February 1994	1	13	2	8
ARA Manager 1	May 1994	2	17	2	8
ARA Manager 2	Jan./Febr. 1994	1	20	1	8
ARA Manager 2	June 1994	2	22	2	7
ARA Manager 3	April 1994	2	20	2	6

Results

Results research question 1: In general, the all in interrater correlation coefficients we found are good or excellent as *average stage scores*. The average rating of participants is higher than the average trainer team rating, self-ratings are somewhat higher than the average participants' rating. However, when taking a look at stage D (Ara Manager Duties) the results are different (see table 2)

Table 2 : Average interrater correlations between different raters on stage scores in the stage ARA Manager Duties

Raters	Seminar				
	AM1 ^a	AM1 ^b	AM2 ^c	AM2 ^d	AM3
Participants/Trainer Team	.1701	.4756	.4119	.4960	.5005
Participants	.1130	.3328	.3488	.4207	.3533
Trainer Team	.4119	.7706	.5366	.8967	.7892

^aFebruary 1994. ^bMay 1994. ^cJanuary 1994. ^dJune 1994. * $p < .01$. ** $p < .001$.

Table 2 shows that the average intercorrelation between participants is unsatisfactory, just like between participants and trainer team. Between the different members of the trainer team the intercorrelations are good in three seminars. On *criterion scores* the average intercorrelations between participants lie in general between .20 and .30. Average intercorrelations between members of the trainer team are at least satisfactory at a number of criteria varying from one to seven. At almost all criteria at least one of the three intercorrelations between members of the trainer team is satisfactory. The conclusion which can be drawn is that only ratings of training professionals are in general satisfactory or almost satisfactory consistent.

Results research question 2:

To determine the consistency between the stages and end score, correlation coefficients are computed. Table 3 shows the results for the stages A, D, E, and F.

In general, intercorrelations between stages and end score are good, but low correlations exist between the Check In and end score of two seminars, and the Check Out of ARA Manager 1 of May and the end score. Analyses of the stages B and C show low correlations of these stages with the end score, except for two of them.

Table 3: Intercorrelations Between Scores in Stages and End Score in the ARA Manager Rallye

Stage	Seminar				
	AM1 ^a	AM1 ^b	AM2 ^c	AM2 ^d	AM3
Check In (A)	.6628*	.2394	.4872	.8138**	.6597**
Check Out (E)	.6814*	.4031	.7915**	.6219*	.6754**
Manager Duty Stage (D)	.6705*	.9203**	.7854*	.8934**	.7539**
Assessment Stage (F)	.7928**	.9321**	.8993**	.8963**	.7841**

^aFebruary 1994. ^bMay 1994. ^cJanuary 1994. ^dJune 1994. * $p < .01$. ** $p < .001$.

Analyses carried out to determine the relationship between the stages show that intercorrelations between stages are low. In three seminars the score of the Check Out correlates lowly with total score and in four seminars with the Check In. In general, scores of the stages B and C correlate lowly, or even negatively, with other stages. The degree to which scores between participants change after each stage in the seminar varies.

Analyses of Stage Scores

If the ratings are determined to be inconsistent between raters, it would be less meaningful to compute an end score consisting of an average of average ratings like the average stage score (Drenth and Sijtsma, 1990). That's why the next step is to determine the interreliability between ratings of participants' and trainer team ratings (research question 1 a), participants (1 b), the ratings of trainer and training professionals (1 c), and between self-ratings and trainer team ratings (1 d)) on the stage score level. The stage score is the sum of the ratings on ten criteria. The results are shown in Table 4 and 5.

Results

The average correlation between participants and trainer team is unsatisfactory. The range between minimum and maximum intercorrelations is relatively large in AM1^a. (In this seminar even more than one third of the correlations is negative). The average intercorrelation between individual participants is unsatisfactory. The range between the minimum and maximum correlation is relatively large in the seminars ARA Manager 1. In the seminars AM1^b, AM2^d, and AM3, good or excellent sizes of average intercorrelations are achieved between members of the trainer team. In these seminars maximum intercorrelations are excellent, minimum correlations good or excellent. In the AM2^c, the correlation coefficient is significant, but does not achieve .60. None of the intercorrelations between self-ratings and trainer team ratings achieves .60, but in the seminar AM3 a significant intercorrelation exist.

Table 4: Average Interrater Correlations Between Different Raters on Stage Scores in the Stage ARA Manager Aufgabe

Raters	Seminar				
	AM1 ^a	AM1 ^b	AM2 ^c	AM2 ^d	AM3
Participants/TT					
Average Correlation	.1701	.4756	.4119	.4960	.5005
Minimum Correlation	-.6718	-.3224	-.3558	-.0518	-.0323
Maximum Correlation	.7377*	.8641**	.7341**	.7962**	.8271**
Participants					
Average Correlation	.1130	.3328	.3488	.4207	.3533
Minimum Correlation	-.7070*	-.6677*	-.4368	-.3008	-.2306
Maximum Correlation	.8461**	.8272**	.8446**	.8167**	.8112**
Trainer Team(TT)					
Average Correlation	.4119	.7706	.5366	.8967	.7892
Minimum Correlation	.2427	.6797*	.5366*	.8756**	.7223**
Maximum Correlation	.6287	.8694**	.5366*	.9261**	.8891**

^aFebruary 1994. ^bMay 1994. ^cJanuary 1994. ^dJune 1994. * $p < .01$. ** $p < .001$.

Table 5: Interrater correlations Between Self-Ratings and Average Ratings of Trainer Team on Stage Scores in the ARA Manager Aufgabe

Stage	Seminar				
	AM1 ^a	AM1 ^b	AM2 ^c	AM2 ^d	AM3
ARA Manager Aufgabe (D)	.1808	.4258	.4909	.3014	.5187*

^aFebruary 1994. ^bMay 1994. ^cJanuary 1994. ^dJune 1994. * $p < .01$

Analyses of Criterium Scores

The stage scores analysed are still scores which result from other scores. They are the sum of ratings on the different criteria of the rating scale used. To be meaningful the sum of ratings on the different criteria has to be based on ratings which are reasonably consistent between the different raters on each criterium. For the seminars ARA Manager 1 of May, and ARA Manager 2 of January, the interrater reliability on criterium scores between participants is determined. For all seminars, the interrater reliabilities between members of the trainer team are determined on every criterium. To determine if there is a difference between the intercorrelations between pairs of members of the trainer team, the average of the intercorrelations on ten criteria is determined for each pair. The results are shown in appendix A.

Results research question 3:

To determine the amount of spread of the end scores of the stages and the end score of the seminar, average scores and their standard deviations are determined in each stage. A total of 1000 points can be gathered. In two seminars the average score is above or almost equals 900 points. In general, average scores lie above 80 points in the stages. In nine of the total of seventeen stages B and C, the average score lies above 900. In the stages in

which only the trainer team judges performances, the spread of scores tend to be much larger than in the stages in which participants determine a part of the end score. Both Check In and Check Out of ARA Manager I of May 1994 have relatively small standard deviations. The average score above 90 in the stage Assessment TL in two seminars is striking, as as the low minimum scores in this stage in four seminars.

Discussion

The results regarding the interrater reliability (research question 1) and the validity (research questions 2 and 3) show that the ratings of participants affect the interrater reliability negatively and that the stages in which they determine a part of the end score show a less spread of scores and high average scores. On the basis of these results a plea can be held to separate the formative and summative stages of the ARAMARK Manager Rallye with regard to the procedure of multirater feedback. In the formative stages B and C, participants should get the chance to improve their performances without much effect on their end scores. These are the stages in which the scores should indicate the progress of performances and in which the participants' learning process can be sharpened. In the summative stages however, assessment of performances should take place by the HRM division management, so that the client determines the quality and not the preferred supplier who delivers the seminars.

To improve the interrater reliability and validity of the multirater test scores a number of improvements are possible:

Since the analyses carried out only considered the scores in the ARA Manager Rallye, more analyses regarding the content of tests, the relationship of the content of tests and the goals of each stage and the content of the stage, are necessary to give exact recommendations regarding the improvement of the interrater reliability and validity of the test scores.

The validity of the test scores of the trainees can be questioned, because 13 out of 22 trainees had responses of over 900 points, which is a very high score, but influences bij trainees themselves.

The use of a Check In to select trainees in the beginning can be questioned too. The group of trainees proves to be not homogeneous which is one of the targets. Moreover, the Check In stage is not very powerful in predicting the end results. One of the reasons for this might be that trainees are tested on knowledge, while most of the seminar is directed on (management) skills. A performance profile could be setup more distinctly, and the criteria could be optimized and tested much better.

Trainers proved to be more consistent and reliable in their ratings, which indicates a more valid and reliable understanding of the assessment criteria involved, the correlations are significantly higher than those of trainees. This is also the case with the assessors of Personnel Department and HRD, with a lesser degree of reliability.

Next suggestions are meant as a tool for optimizing the Rallye as a training and assessment tool:

- a) separation of formative and summative evaluation;
- b) controlling of the real learning effects in Check In and Check out: not only ratings of colleagues, but also performance measurements;
- c) changing the rating scale, which is mixture of nominal and interval scale;
- d) optimizing the (number and the content of) test items, which proved to be not reliable;
- e) a better training procedure for trainees to assess on the basis of well operationalized test items.

When the rallye has to function as a really good training tool in skills, the first thing is to have a reliable rating instrument, which seems to be not very easy when having a look in literature. Personnel decisions and career development based on such an instrument have to be reconsidered or at least to be completed with other valid job information.

The assumptions behind the instrument are well elaborated, but practice is harsh.

References

- Bracken, D.W. (1994). Straight Talk about Multirater Feedback. *Training & Development*. September, (pp. 44-51),
- Bell, C. & Harris, Ch. (Eds.) (1990). *Assessment and Evaluation*. New York: Kogan Page.
- Drenth, P. & Sijtsma, K. (1990). *Testtheorie*. Inleiding in de theorie van de psychologische test en zijn toepassingen. Houten: Van Loghum Slaterus. [Test Theory: Introduction into the theory of the psychological test and its applications].
- Russel. T. (1994). *Effective Feedback Skills*. London: Kogan Page.

Appendix A: Average Interrater correlations Between Participants and Between Trainer Team in the ARA Manager Aufgabe

Raters	Criterium									
	1	2	3	4	5	6	7	8	9	10
ARA Manager 1 of May										
Participants	.2530	.2571	.2503	.1958	.0336	.1385	.1000	.1728	.1799	.2397
Trainer Team	.5639	.5690	.7280	.6560	.5223	.3979	.3670	.5610	.6944	.4003
Minimum Correlation	.4133	.4821	.6880*	.5277	.3052	.2890	.3069	.2997	.6429*	.3179
Maximum Correlation	.6832*	.7275*	.7242**	.8200**	.7048**	.6215*	.4336	.7015*	.7694**	.4578
ARA Manager 2 of January										
Participants	.2045	.2471	.3025	.1103	.1791	.0865	.1369	.2346	.3266	.2687
Trainer Team	.5648	.3388	.5977	.3790	.4658	.1591	.4737	-.0803	.7209	.5053
ARA Manager 2 of June										
Trainer Team	.7798	.6435	.7234	.5938	.5169	.6203	.4216	.6898	.7162	.7340
Minimum Correlation	.7393*	.5840*	.5790*	.5006*	.3839	.5962*	.2877	.4926*	.6704**	.6815**
Maximum Correlation	.8146*	.6950*	.9177**	.7691*	.6146*	.6664*	.5820*	.8002**	.7420**	.8037**
ARA Manager 3										
Trainer Team	.7129	.7460	.3083	.3862	.3552	.5186	.6683	.4511	.3617	.3798
Minimum Correlation	.6549*	.6535*	.1737	.1426	.2663	.3498	.5493*	.3677	.2335	.1132
Maximum Correlation	.7505*	.8960*	.3793	.7148*	.4889	.7547**	.8812**	.5248*	.4958	.5250*

In Search of an Integrative Model for HRD Evaluation

Elwood F. Holton III
Louisiana State University

The lack of research to further develop a theory of evaluation is a glaring shortcoming for HRD. In this paper, it is argued that the four-level Kirkpatrick system of training evaluation is really a taxonomy of outcomes and that research is needed to develop an integrated measurement model. Such a model needs to account for the effects of intervening variables that affect training outcomes and specify causal relationships. An initial integrative evaluation model that accounts for the impact of the primary intervening variables such as motivation to learn, trainability, job attitudes, personal characteristics and transfer of training climate is proposed. A new role for trainee reactions is specified. Key studies supporting the model are reviewed and a research agenda is proposed.

Evaluation of interventions is one of the most critical issues facing the field of Human Resource Development (HRD) today. Increasing global competition has led to intense pressure on HRD to demonstrate that programs are directly contributing to the "bottom line" of the organization. Yet, the pre-eminent evaluation model, the four level Kirkpatrick model, has received alarmingly little research and is seldom fully implemented in organizations (Kimmerling, 1993) leaving them ill-equipped to respond to this pressure. A critical need exists for sound research on this widely accepted evaluation model to give organizations a more sound methodology for allocating HRD resources.

The Kirkpatrick model (Kirkpatrick, 1976) for training evaluation, also known as "4-level evaluation," is widely accepted by practitioners as the standard in the field. A number of modifications to the model have been suggested, including adding a fifth level to reflect training's ultimate value (Hamblin, 1974), societal value (Kaufman & Keller, 1994) or to more specifically focus on ROI (Phillips, 1995). Brinkerhoff (1987) proposes a six level model that, in essence, adds two formative evaluation stages as precursors to Kirkpatrick's four levels. While this work has contributed greatly to our conceptual thinking about evaluation, the models have received only limited empirical testing. All of them are best labeled as taxonomies which are simply classification schemes (Bobko & Russell, 1991).

Bobko and Russell (1991), citing Wallace (1983), noted that exploratory designs and case studies are the first steps in the development of theory, while the final steps are correlational and experimental studies. According to them, taxonomies are the link between the initial stages and the final confirmatory stages of developing theory. While the Kirkpatrick model is elegant in its simplicity and is a very useful taxonomy, the lack of research to further develop a theory of evaluation is a glaring shortcoming for HRD. If HRD is to continue to grow as a profession, an evaluation model grounded in research is necessary.

One shortcoming of taxonomies is that they do not fully identify all the constructs underlying the phenomena of interest, making confirmatory evaluation impossible. Not surprisingly then, Alliger and Janak (1989), in their comprehensive review of the research on the Kirkpatrick model, note that the implied causal relationships between each level of this somewhat simple taxonomy have not been demonstrated by research. Their search of the relevant academic literature located only 12 articles since 1959 reporting 26 correlations between levels in training programs out of 203 articles that reported any type of evaluation results. Furthermore, only three studies (Clement, 1982; Noe & Schmitt, 1986; Wexley & Baldwin, 1986) reported full 4-level evaluations with correlations. The reported correlations varied widely, casting doubt on these assumptions and providing very limited support for the causal relationships.

It is not surprising that the correlations are weak since the "model" is really a taxonomy of training (and HRD) outcomes. Attempts to test causal assumptions within a taxonomy are futile because,

by definition, taxonomies just classify, not define causal constructs. In fact, it can be argued that the correlations reported in these studies are not really a test of the model but instead an alternate approach to analyzing outcomes of the training. For example, if only the four levels of outcomes are measured and a weak correlation is reported between levels 2 and 3, all we really know is that learning from training was not associated with behavior change. In the absence of a fully specified model, we don't really know if the correlation is weak because some aspect of the training effort was not effective, or because the underlying evaluation model is not valid. Weak correlations might represent a well functioning model reporting a poorly functioning training effort.

It can be argued then that the Kirkpatrick model is inadequate as a measurement and research model. Klimoski (1991), citing Dubin (1976), notes that theories or models should have six components:

1. *Elements or units*, represented as constructs, that are the subject matter.
2. *Relationships* between the units.
3. *Boundaries or limits* of generalization.
4. *System states* and changes described.
5. *Deductions* about the theory in operation, expressed as propositions or hypotheses.
6. *Predictions* about units (p. 254-256).

The Kirkpatrick model is inadequate when assessed against all six of these criteria. Noticeably absent from the Kirkpatrick model are the large number of intervening variables that affect learning and transfer processes such as trainee readiness and motivation, training design, and reinforcement of training on the job (Clement, 1982). Others have proposed models of how individual differences affect training outcomes (Noe, 1986; Noe & Schmitt, 1986) and factors affecting the transfer of training (Baldwin & Ford, 1988; Broad & Newstrom, 1992). Previous evaluation studies identified by Alliger and Janak (1989) did not attempt to measure any intervening variables which is one likely reason for the wide variation in the relationships reported. It seems clear that no evaluation model can be validated without measuring and accounting for the effects of intervening variables.

A critical need exists for intensive research to move from a taxonomic approach to evaluation, to a fully specified model that meets the six criteria of good theories and models outlined above. The purpose of this paper is to take an initial step toward establishing a fully validated model for HRD evaluation. An integrative evaluation model that accounts for the impact of the primary intervening variables such as trainability and transfer of training climate is proposed. Key studies supporting this model are reviewed. It is hoped that this paper, and the dialogue that follows, will spark an intensive search for a more integrative and testable model of HRD evaluation.

Research Supporting Integrative Evaluation Models

Two studies have demonstrated the importance and feasibility of this approach by examining influences on training outcomes. Noe (1986) used path analysis to examine the relationship of trainee reactions to skill assessment, job involvement, career planning, exploratory behavior, locus of control, post-training motivation and pre-training motivation on reactions, learning, behavior change and performance improvement. Mathieu, Tannenbaum & Salas (1992) used a more powerful technique, structural equation modeling, to examine the relationships between situational constraints, assignment to training, motivation, previous education, trainee reactions, learning and behavior change.

Specific results of these studies will be discussed below, but several general conclusions can be noted that serve to justify this line of research. First, both studies supported Alliger & Janak's (1989) suggestion that "reactions" are not linearly related to learning, but instead may act as a moderator of learning. Neither found support for reactions as a primary outcome of training. Second, both studies found that complex relationships exist between the various intervening variables and the primary outcomes of learning and performance change. Third, they suggest that further research holds promise for developing a fully integrated model.

These studies also suggest that the four-level evaluation model is woefully inadequate as a diagnosis tool. Consider the case where learning is not found to occur. The usual conclusion is that

something is wrong with the training program. However, if one considers the many intervening variables that remain unmeasured, it is quite possible that the training program is well-designed and the problem lies outside the classroom with some element of the job or the individual. With a fully specified measurement model, the true problems with the training could be isolated. Furthermore, through proper statistical analysis that controls for the effects of intervening variables, it might be possible to show an effective training program design even when the overall group learning scores indicate poor learning outcomes. This is not to suggest that trainers are only responsible for what occurs in the training room, because in this example, if the outcomes are not positive then trainers have not accomplished their goals. Rather, it suggests that there is a complex system of influences on training outcomes that must be measured if training is to be accurately evaluated. This model would enable practitioners to correctly diagnose barriers to training effectiveness.

Proposed Model

The proposed evaluation model is shown in Figure 1. Three primary *outcome* measures (learning, behavior change and results) are shown at the right side of the model. *Primary intervening variables* (ability, motivation to learn, reaction to learning, transfer design, motivation to transfer, and external events) are shown in boxes and have arrows pointing directly to one of the outcomes. *Secondary intervening variables* (trainee readiness, job attitudes, personality characteristics, transfer climate) do not have boxes around them and have arrows pointing directly to one of the primary intervening variables.

The model as presented here describes a sequence of influences on outcomes occurring in a single learning experience. Over time, it would be expected that employees' successes at achieving results after learning experiences would enhance future motivation to learn. Said differently, over time there are many cumulative feedback loops that are not shown in detail in this model. A single feedback loop is used to represent the recurring nature of this process and the cumulative effects on the learner over time. The model itself is limited to evaluating outcomes from a single learning intervention.

Influences on Learning Outcomes

This model posits that there are three primary influences on learning: trainee reactions, motivation to learn, and ability.

The Role of Trainee Reactions Perhaps the most discussed portion of the Kirkpatrick model is the inclusion of reactions as the first level. Dixon (1990) demonstrated that there is little correlation between reactions and learning. Indeed, most learners would acknowledge that good learning can often be confusing and frustrating. While some may argue that trainee reactions are unimportant, as a practical matter few practitioners can afford to totally ignore the reaction of their trainees. Cognitive scientists point out that trainee reactions can play a role in building interest and attention and enhancing motivation.

The question here is whether trainee reactions are a primary outcome of training that is linearly related to learning, or simply another intervening variable that impacts upon learning. Mathieu et al. (1992) found that reactions functioned as a moderator of the relationship between training motivation and learning as well as a mediator of other relationships. These results offer the first explanation for the generally low correlations between reactions and learning observed by Aliger and Janak (1989). The author's note that had they examined only the linear relationship, they too would have concluded that reactions had no significant effect on learning. Instead, they conclude that they are important, but not in and of themselves. Similarly, Noe (1986) found no support for a direct link between reactions and learning.

While these findings fly in the face of many years of training practice, and may challenge the humanistic values of many practitioners, they suggest that trainee reactions should be removed from the model as a primary outcome of training. Rather, they are included only as having a complex role moderating the relationship between motivation to learn and learning. Trainee reactions are posited to

Figure 1 - Evaluation Measurement Model

```
graph TD; EE[External Events] --> R[4 - RESULTS]; TD[Transfer Design] --> B[3 - BEHAVIOR]; MT[Motivation to Transfer] --> B; TC[Transfer Climate] --> MT; RCL[Reaction] <--> L[2 - LEARNING]; RL[Motivation to Learn] --> RCL; RL --> L; RL --> B; RL --> R; RL --> A[Ability]; P[Personality Char.] --> RL; JA[Job Attitudes] --> RL; TR[Training Readiness] --> RL; A --> L; L --> B; B --> R; R --> FL[Feedback Loop]; FL --> EE; FL --> TD; FL --> TC; FL --> P; FL --> JA; FL --> TR; FL --> A;
```

The diagram illustrates the Evaluation Measurement Model, showing the progression from external events to results, influenced by various factors like motivation, learning, and ability.

Key Components and Flow:

- External Events** leads to **4 - RESULTS**.
- Transfer Design** leads to **3 - BEHAVIOR**.
- Motivation to Transfer** leads to **3 - BEHAVIOR**.
- Transfer Climate** leads to **Motivation to Transfer**.
- Reaction** and **2 - LEARNING** are interconnected by a double-headed arrow.
- Motivation to Learn** leads to **Reaction**, **2 - LEARNING**, **3 - BEHAVIOR**, and **4 - RESULTS**.
- Personality Char.** and **Job Attitudes** lead to **Motivation to Learn**.
- Training Readiness** leads to **Motivation to Learn**.
- Ability** leads to **2 - LEARNING**.
- 2 - LEARNING** leads to **3 - BEHAVIOR**.
- 3 - BEHAVIOR** leads to **4 - RESULTS**.
- A **Feedback Loop** connects the final results back to the initial conditions (External Events, Transfer Design, Transfer Climate, Personality Char., Job Attitudes, Training Readiness, and Ability).

have a reciprocal relationship with learning, where more positive reactions to the training are expected to enhance learning and trainees who are more successful during learning are expected to have more positive reactions to the learning experience.

Pre-training Motivation Trainee motivation to learn then has a direct relationship with learning. Pre-training motivation and trainee attitudes have not received enough attention in the literature. Three categories of variables are hypothesized to be primary influences on motivation to learn: trainee readiness for training, job attitudes, and personality characteristics.

Trainee Readiness. Several studies have examined influences on trainees' readiness to enter and participate in training programs. Hicks and Klimoski (1987) found that giving trainees a choice of whether to attend training or not increased motivation to learn and improved learning outcomes. Baldwin, Magjuka and Loher (1991) further found that trainees having a choice of training content had greater motivation to learn. Those who were allowed to choose, but then were not given their choice of training, became less motivated than those who were not allowed to choose at all. Thus, the degree to which a trainee is involved in the needs assessment process and given choices about training would be expected to influence motivation to learn. The notion of meeting trainees expectations and desires for training has also received support in the literature (Tannenbaum, Mathieu, Salas, & Cannon-Bowers, 1991; Hicks & Klimoski, 1987). Those trainees who do not feel the training will meet their needs will be less motivated and less likely to learn.

It is likely that motivation to learn will vary by trainee's readiness for training. Readiness for training includes such variables as the degree to which trainees are involved in assessing needs, involvement in planning the training, degree to which expectations are clarified, degree of choice, and other influences.

Job attitudes. An unexplored influence on motivation to learn is the attitude of the trainee toward their organization and job. Only two studies could be located that tested this notion. Noe (1986) found a significant relationship between job involvement and learning while Tannenbaum et al. (1991) found that more committed employees performed better in training. However, Mathieu et al. (1992) did not find a significant relationship between job involvement and motivation, but attributed it to the type of training in the study. It is likely employees who exhibit more positive job attitudes would be more motivated to learn and, in turn, have more positive training outcomes. Job attitudes are hypothesized to have an indirect influence, working in combination with other factors.

Personality characteristics. Selection researchers have long been interested in the validity of personality measures in predicting performance. The "Big Five" personality dimensions (extroversion, openness to experience, neuroticism, agreeableness and conscientiousness) have been shown in a recent meta-analysis to have validity in explaining some of the variance in performance (Tett, Jackson & Rothstein, 1991). While HRD researchers have not explored these directly, other characteristics such as self-efficacy (Gist, Stevens & Bavetta, 1991) and locus of control (Noe, 1986) have been shown to be related to training outcomes. Thus, certain personality characteristics would be expected to influence motivation to learn and, in turn, learning.

Ability Long forgotten in the evaluation literature is the role general cognitive ability plays in influencing training outcomes. When evaluating a training program with a relatively homogenous group of trainees in terms of job level and educational background, it is quite likely that there is little variance in general ability and therefore little detectable influence on training outcomes. In situations where trainee groups are heterogenous in cognitive ability, or where programs are offered across an organization to groups of varying ability, it is likely that general cognitive ability would influence training outcomes. Because it is difficult, if not impossible, to control for ability through random samples in most evaluation studies, it is essential to measure and control for it statistically. Trainees with greater ability will be expected to perform better in learning activities.

Influences on Job Behavior Outcomes

Two primary influences on transfer behavior are proposed: motivation to transfer and transfer design.

Motivation to Transfer Trainees leave a training program with a certain level of motivation to utilize the training on the job. Traditional evaluation models acknowledge that transfer to the job is not

certain, but fail to measure influences on transfer. A variety of influences have been suggested (Broad & Newstrom, 1992; Baldwin & Ford, 1988) that fall into two primary categories: transfer climate and reaction to training.

Transfer climate. The notion that situational constraints affect workplace performance has become established in the literature (Peters, O'Conner & Eulberg, 1985). Less certain is which constraints directly affect the transfer of training. Recently, Rouillier and Goldstein (1993) reported on the development of an instrument to measure transfer climate. They proposed that transfer climate consisted of seven dimensions: goal cues, social cues, task cues, positive reinforcement, negative reinforcement, punishment, and extinction. While they were not successful in validating the separate subscales in the instrument, their promising work demonstrated that transfer climate added significantly to the variance in performance explained after controlling for learning and unit performance. Tracy (1992) attempted to replicate their work, using their instrument along with his own organizational learning measures. He also found that transfer climate explained a significant portion of the variance in transfer. In addition, he refined their instrument and validated the use of an organizational learning climate scale.

Research to develop instruments to measure transfer is still in its infancy and much work remains to be done. However, even while using developmental instruments, this research has clearly shown what has been suggested in the practitioner literature for some time: non-training factors such as supervisor support for training and rewards for using training affect a trainee's motivation to transfer learning into behavior change on the job. Without controlling for the influence of transfer climate, evaluation results are likely to vary considerably and yield erroneous conclusions about the causes.

In this model, transfer climate is posited to affect motivation to transfer which in turn influences transfer of training. Trainees who work in a climate supportive of training are more likely to transfer their learning to the job. In addition, transfer climate influences motivation to learn because these trainees would have greater motivation to learn, as well as transfer.

Reaction to training. Trainee reactions are posited to have a secondary influence on motivation to transfer. It seems reasonable that trainees who had positive reactions to training would be more motivated to transfer training to the job than would trainees who had negative reactions. Tannenbaum et al. (1991) found a correlation of .45 between reactions and post-training motivation. In hierarchical regression analyses, reactions were found to be positively related to post-training commitment and motivation. Tziner, Haccoun and Kadish (1991) reported a .54 correlation between reactions and motivation to transfer. Similarly, Mathieu et al. (1992) found a significant relationship between reaction to training and post-training performance, though they did not obtain a motivation to transfer measure.

In this model, reactions are expected to impact upon motivation to transfer and work in combination with transfer climate to influence transfer. Trainees who have a positive experience in training are expected to be more motivated to use the training on the job.

Transfer Design Another likely cause of failure to transfer is that the design of the training does not provide trainees the ability to transfer the learning. That is, cognitive learning may well occur but trainees may not have an opportunity to practice the training in a job context or not be taught how to apply their new knowledge on the job. For example, several studies (Werner, O'Leary-Kelly, Baldwin, Wesley, 1994; Wexley & Baldwin, 1986) have shown that goal setting during and after training improves transfer.

In a more complex task situation such as negotiation training, improved results were obtained by augmenting goal setting with self-management training (Gist, Bavetta, & Stevens, 1990). Tziner et al. (1991) showed that adding a relapse prevention module to training resulted in higher learning and greater transfer.

The degree to which transfer mechanisms are included in the design of the training itself is hypothesized to have a direct influence on the transfer of training. Trainees who are taught how to apply new knowledge and skills in a job context should have the ability to transfer the learning which, when combined with the motivation to transfer, is likely to result in greater transfer.

This component of the model is probably the most difficult to measure because there are few definitive guidelines for what constitutes appropriate transfer designs. Transfer designs are expected to vary considerably depending on different content, cultures and other situational factors. Nonetheless, methods must be developed to assess the extent to which trainees acquire the ability to transfer the learning. Clearly even the most motivated trainee will be unable to transfer the learning without knowing how to do it.

Influences on Results

One of the biggest challenges in evaluating organizational results (level 4) is isolating the effects of training. A wide variety of factors can impact upon organizational results that are completely outside the realm of training. For example, in a manufacturing environment productivity might be influenced by equipment failures, raw material shortages, price changes, absenteeism, etc. The key external factors need to be identified and controlled for. A variety of approaches can be used, ranging from simple control groups to more complex statistical methods. However, because the strategies are unique to each individual organization and training program, they can not be specified in this model beyond a single global category of external factors.

Validating the Model

The validation of this model will clearly be an ambitious undertaking. One difficulty is in finding training sites that are willing to donate the time to collect all this data. Testing of this model will require the use of sophisticated statistical techniques; the most promising of which is structural equation modeling (e.g. LISREL) which can evaluate all the causal relationships in the model simultaneously. However, this technique requires relatively large samples that are often difficult to obtain. Neither of these challenges are insurmountable. In fact, this researcher has several projects underway that should produce data to validate a significant portion of the model.

From this discussion though it should be evident that this is an effort that needs to be undertaken. From a research perspective, the simple four level taxonomy is totally inadequate. As suggested above, it is likely that the role of reactions is much different than that specified in the Kirkpatrick model. Furthermore, the model can never be validated because there are too many intervening variables that are unmeasured. It should be recognized that it is only a taxonomy, not a research model, and researchers should cease trying to validate it in its present form.

It should be equally evident though that the development and validation of a more complete model is quite feasible. Recent research points to the vital components of such a model and to the methodologies for measuring and validating it. The research agenda to move toward an integrative model should include the following:

- Validating the basic components of an integrative model.
- Identifying the relationships within the model, including classifying direct, moderating and mediating effects.
- Identifying the specific variables that should be measured within each of the major components of the model
- Developing and testing methodologies to analyze such a complex model.
- Demonstrating that a fully integrated model can explain a major portion of the variance in training outcomes.

There are practical implications for this model as well. Practitioners who use only the four level model are quite likely to arrive at erroneous conclusions about the success of their training programs. They need a model that can be used as a diagnostic tool, directing them to critical influences that need to be measured along with the outcomes and lead to accurate explanations of the outcomes obtained. It is possible that, after further research, a simpler model could be developed for routine application where much of the variance can be explained by a smaller set of variables. For instance, it may be that measuring the primary intervening variables alone will be sufficient. Or, there may be a few key variables within each category that would be sufficient to measure. There is no theoretical reason that a practitioner model more sophisticated than the Kirkpatrick model, but simpler to implement than the one described here, could not be developed.

It is ironic that this model may point the way to a more important role for reaction forms in a redesigned format. Many of the arguments about reaction forms are really the result of failing to properly specify their relationship to the more important outcomes of learning and performance. Those who argue against the use of reaction forms are correct in that they should not be a primary outcome. Those who

argue for their use are also correct because they intuitively recognize that reactions do affect outcomes. The four-level taxonomy offers no other option; this model puts reactions in the proper category of an important influence on outcomes, but not an outcome in and of itself. An interesting research question is whether the reaction forms that collect self-reports of motivation to transfer, transfer design, motivation to learn, personal characteristics, transfer climate, etc. might be reasonable surrogate measures of these constructs. If so, then the measurement process could be greatly simplified for practitioners.

Conclusion

It seems that thirty-five years is much too long to rely on such a simple and taxonomy of training evaluation. In this paper I have argued for moving away from a taxonomic evaluation system to the development and testing of a true model of HRD evaluation. The model presented here is an initial step in that direction and will surely be refined as research is conducted to test it. It should be realized from looking at this model that the development and validation of a fully specified HRD evaluation model is within our reach. The methodologies to analyze such a model are now well established and sufficient research exists to develop reasonable research hypotheses. Reliance on the simple four level taxonomy only serves to minimize the value, impact and sophistication of the intervention tools we employ. If HRD is to grow as a discipline and as a profession, it seems imperative that researchers work deliberately to develop a more integrative and testable model.

References

- Alliger, G. M. & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personal Psychology*, 42, 331-340.
- Baldwin, T. T. & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology*, 41, 63-100.
- Baldwin, T. T., Magjuka, R. J., & Loher, B. T. (1991). The perils of participation: Effects of choice of training on trainee motivation and learning. *Personnel Psychology*, 44, 51-65.
- Bobko, P. & Russell, C. (1991) A review of the role of taxonomies in human resources management. *Human Resource Management Review*, 4, 293-316.
- Brinkerhoff, R. O. (1987). *Achieving results from training*. San Francisco: Jossey-Bass.
- Broad, M. L. & Newstrom, J. W. (1992). *Transfer of training*. Reading, MA: Addison-Wesley Publishing Company, Inc.
- Carnevale, A. P. & Schulz, E. R. (1990, July). Return on investment: Accounting for training. *Training and Development*, 23-29.
- Clement, R. W. (1982). Testing the hierarchy theory of training evaluation: An expanded role for trainee reactions. *Public Personnel Management Journal*, 11, 176-184.
- Dixon, N. M. (1990). The relationship between trainee responses on participation reaction forms and posttest scores. *Human Resource Development Quarterly*, 1, 129-137.
- Dubin, R. (1976). Theory building in applied areas. In M. D. Dunnette (Ed.), *Handbook of Industrial/Organizational Psychology*. New York: Rand McNally.
- Gist, M. E., Bavetta, A. G., & Stevens, C. K. (1990). Transfer training method: Its influence on skill generalization, skill repetition, and performance level. *Personnel Psychology*, 43, 501-523.
- Gist, M. E., Stevens, C. K. & Bavetta, A. G. (1991). Effects of self-efficacy and post-training intervention on the acquisition and maintenance of complex interpersonal skills. *Personnel Psychology*, 44, 837-861.
- Hamblin, A. C. (1974). *Evaluation and control of training*. London: McGraw-Hill.
- Hicks, W. D. & Klimoski, R. J. (1987). Entry into training programs and its effects on training outcomes: A field experiment. *Academy of Management Journal*, 30, 542-552.
- Kaufman, R. & Keller, J. M. (1994). Levels of evaluation: Beyond Kirkpatrick. *Human Resource Development Quarterly*, 5, 371-380.

- Kimmerling, G. (1993, Sept). Gathering the best practices. *Training and Development*, 29-36.
- Kirkpatrick, D. L. (1976). Evaluation of training. In R. L. Craig (Ed.), *Training and development handbook*, (pp. 18-1 - 18-27). New York: McGraw-Hill.
- Klimoski, R. (1991). Theory presentation in human resource management. *Human Resource Management Review*, 4, 253-271.
- Noe, R. A. & Schmitt, N. (1986). The influence of trainee attitudes on training effectiveness: Test of a model. *Personnel Psychology*, 39, 497-523.
- Noe, R. A. (1986). Trainee attributes and attitudes: Neglected influences on training effectiveness. *Academy of Management Review*, 11, 736-749.
- Mathieu, J. E., Tannenbaum, S. I., & Salas, E. (1992). Influences on individual and situational characteristics on measures of training effectiveness. *Academy of Management Journal*, 35, 828-847.
- Peters, L. H., O'Conner, E. J., Eulberg, J. R. (1985) Situational constraints: Sources, consequences, and future considerations. In B. Rowland and G. Ferris (Eds.) *Research in personnel and human resources management*, vol. 3, (pp. 79-113). Greenwich, CN: JAI Press.
- Phillips, J. (1995). Return on investment -- Beyond the four levels. In E. Holton (Ed.) *Academy of HRD 1995 Conference Proceedings*.
- Rouillier, J. Z. & Goldstein, I. L. (1993) The relationship between organizational transfer climate and positive transfer of training. *Human Resource Development Quarterly*, 4, 377-390.
- Tannenbaum, S. I., Mathieu, J. E., Salas, E. & Cannon-Bowers, J. A. (1991). Meeting trainee's expectations: The influence on the development of commitment, self-efficacy, and motivation. *Journal of Applied Psychology*, 76, 759-769.
- Tett, R. P., Jackson, D. N., & Rothstein, M. (1991) Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, 44, 703-742.
- Tracy, J. B. (1992). *The effects of organizational climate and culture on the transfer of training*. Unpublished doctoral dissertation, State University of New York at Albany.
- Tziner, A., Haccoun, R. R., & Kadish, A. (1991). Personal and situational characteristics influencing the effectiveness of transfer of training improvement strategies. *Journal of Occupational Psychology*, 64, 167-177.
- Wallace, M. (1983). Methodology, research practice, and progress in personnel and industrial relations. *Academy of Management Review*, 8, 6-13.
- Werner, J. M., O'Leary-Kelly, A. M., Baldwin, T. T., Wexley, K. N. (1994) Augmenting behavior modeling training: Testing the effects of pre- and post-training interventions. *Human Resource Development Quarterly*, 5, 169-183.
- Wexley, K. N. & Baldwin, T. T. (1986). Post-training strategies for facilitating positive transfer: An empirical exploration. *Academy of Management Journal*, 29, 503-520.

Building Training on the Pillars: Applying Total Quality in Human Resource Development

Douglas H. Smith and Jo D. Gallagher
Florida International University

The application of total quality management in human resource development is the focus of this paper. An overview of five leading total quality writers is summarized, from which four basic principles, or pillars, are derived and serve as the basis of the application of total quality with the instructional setting. Examples of the relationship of the four pillars to instructional delivery are then presented. Comments regarding instructional evaluation, and the need to expand the current framework of formative and summative evaluation to include implementative evaluation concludes the paper.

A major world wide force in organizational development, public and private, is the total quality movement. Launched initially in Japan, TQM and CQI (continuous quality improvement) is becoming an integral part of the lexicon of organizational literature. Organizations are rapidly moving to implement total quality practices, with many attempts being highly successful, but too many are becoming dismal failures.

A major chapter in many TQM/CQI success stories is the integration of education and training within the total quality functions of the organization. Much has been written of the need and importance of the human resource development program playing a central role in any TQM/CQI effort. What has not been emphasized as much is the need to assimilate within the HRD program the principles and practices of TQM/CQI. The instruction itself should reflect and demonstrate the demeanor or cultural behavior desired within the organization.

This paper addresses the issue of applying TQM/CAI in HRD. It begins first with an overview of the major contributions of leading total quality writers, from which four basic principles, or pillars, are derived and serve as the basis of the application of total quality with the instructional setting. Examples are given to illustrate the relationship of the four pillars to instructional delivery. The paper concludes with an observation regarding instructional evaluation, and the need to expand the current framework of formative and summative evaluation to include another central function, identified as implementative evaluation.

The Basis of TQM/CQI

The five recognized gurus of the total quality management are W. Edwards Deming, Armand A. Feigenbaum, Joseph M. Juran, Kaoru Ishikawa, and Philip B. Crosby. Individually they have molded the framework upon which present TQM/CQI practices are based. Table 1, on the following two pages, presents the major functions described by each writer within six sequential steps, from the initial commitment to total quality through to the solidification of the gains achieved.

From an analysis of the commonalities of the basic philosophies and operational concepts of these five writers, four common principles can be identified and described. Together they form the basis of both the application of TQM/CQI within the organization, and the integration of TQM/CQI in the instructional, HRD, development and delivery. The principles, identified by the authors as four pillars, are discussed (following Table 1) within the context of their organizational relationship and their application in HRD.

Table 1. Quality Improvement Concepts - A Comparison of W. Edwards Deming, Armand A. Feigenbaum, Joseph M. Juran, Kaoru Ishikawa/J.U.S.E., and Phillip B. Crosby¹

THE COMMON PHASES/STEPS	THE DEMMING MODEL	THE FEIGENBAUM/JURAN MODEL	THE ISHIKAWA/J.U.S.E. MODEL	THE CROSBY MODEL
1. Steering Arm	1 Create constancy of purpose.	1 Create the steering arm.	1 Create the design and development team.	1 Commitment from management.
	2 Adopt a new philosophy.	2 Define the scope, charter.	2 Define the scope, charter.	2 Communicate Quality Awareness.
	7 Institute leadership.	3 Appoint members and train.	3 Appoint members and train.	3 Create Quality Improvement Organization.
		4 Create the internal machinery.	4 Create the machinery.	4 Form a Quality Council.
	13 Education and retraining.	5 Form a company-wide task force to ensure breakthrough.	5 Form a Quality Council.	5 Measure non-conformance
	5 Improve constantly the system.	6 Design the diagnostic package and instruments.	6 Create management review structures.	
	8 Drive out fear.	7 Collect data and analyze prior operations.	7 Perform annual review of operations problems.	
	9 Break down dept. barriers.	8 Create new data, collect and assemble.	8 Conduct customer market survey.	
		9 Analyze data and interpret.	9 Analyze data and interpret results.	6 Set goals.
2. Diagnosis Journey		10 Create a mosaic of knowledge to serve as a guide to management decision and action.	10 Create the policy deployment package, vision, objectives and tactical items as a guide to management decision and action.	7 Plan Quality Day.
	6 Training and retraining.	11 Train pilot areas and assign priority to projects.	11 Train pilot areas, select projects and develop action plans.	8 Quality Education.
	4 /end Price tags awards.	12 Summarize past data; conduct pareto analysis of symptoms.	12 Gather specific historical data relating to each project.	9 Cost of Quality Evaluation.
		13 Theorize on causes of symptoms.	13 Analyze data, develop decision themes and lay out structures.	
		14 Test theories; collect and analyze data on capabilities.	14 Develop problems specs and identify additional data needed.	
		15 Narrow list of theories; identify experimental data needed.	15 Present project data to lead team using the QI Story and obtain approval.	
		16 Design experiments.	16 Analyze for possible root causes, using fishbone, pareto, structure free, etc.	
		17 Approve design; provide authority.		
		18 Conduct experiment; collect, analyze and present data; embellish proof of cause.		
3. Project Journey - From Symptom to Cause				

Table 1. Quality Improvement Concepts - A Comparison of W. Edwards Deming, Armand A. Feigenbaum, Joseph M. Juran, Kaoru Ishikawa/J.U.S.E., and Phillip B. Crosby (continued)¹

THE COMMON PHASES/STEPS	THE DEMMING MODEL	THE FEIGENBAUM/JURAN MODEL	THE ISHIKAWA/J.U.S.E. MODEL	THE CROSBY MODEL
4 Project Journey - From Cause to Remedy	6 Training and retraining	19 Consider and propose remedies.	17 Select possible solutions and remedies.	
	10 Eliminate slogans and targets.	20 Test remedy.	18 Perform trial implementation.	
		21 Action to institute remedy; control at new level.	19 Take action to implement solution.	
5 Remedial Journey	6 Training and retraining	22 Install tools for early warning signs and monitoring.	20 Create indicators and evaluate impact.	
	11 Eliminate numerical quotas.	23 Design and implement methods of foolproofing the "hardware" and communications.	21 Design feedback mechanisms and standardize terms and glossaries.	11 Remove cause of error.
	3 Cease dependence on mass inspection.	24 Deal with resistance to change through direction, training and recognition.	22 Assure involvement of key managers through policy deployment and total recognition.	12 Recognition
	12 Remove barriers to workman's pride.	25 Develop detailed specifications to assure fitness for use.	23 Develop valid requirements based upon customer needs and wants.	13 Quality Day
6. Holding the Gains	6 Training and retraining.	26 Designing the control process.	24 Develop a control system for implemented policy deployment project, outline, and test.	
	4 /end Price tags awards.	27 Provide the capable process.	25 Develop new operating procedures and guidelines.	
		28 Establish new operating standards and procedures.	26 Train the operating forces.	8 Quality Education.
		29 Train the operating forces.		

¹Adapted from Voehl, Frank W. (1992). *Total Quality Printiples and Practices Within Organizations* (pp. 16 b,c,d)

Customer satisfaction

As illustrated in Figure 1, one of the few points that all TQ writers agree is that the core purpose of any quality improvement process is to ensure that the needs and reasonable expectations of the customer are satisfied. Any organization, public or private, service or manufacturing, will prosper when this occurs. Serving the customer is straightforward, i.e., identifying the customer (Who are our customers?), communicating with the customer and agreeing on their needs and reasonable expectations that are both desirable and possible (What do they want?), and designing the work process to be sure these requirements are met (What must we do to meet and even anticipate our customers' needs?).

To meet the expectations of the customer, many writers identify an *external customer* (the ultimate recipient of the organization's product or service), and an *internal customer* who is the next recipient of each unit's or team's service or product. Thus, while many do not directly serve the external customer, the attention to service to the unit's being served ultimately assures high quality productivity throughout the organization.

Continuing within the context of multiple customers, HRD has a number of customers for two basic services of HRD, instructional development and organizational development (Table 2). The *internal customers of instructional development* are the learners (teams and individual employees) needing appropriate knowledge, skills and abilities; departments needing continuous improvement and information; and instructors and/or consultants needing continuous learning both the product or service of the organization, and instructional delivery skills. The *internal customers of organizational development* are the executive officers who need consultation on the maximum utilization of organization's human resources, i.e. the HRD plan, and participation on the Quality Council; and the units of the organization who need assistance in the implementation of the HRD plan. Two external customers may exist for HRD: *Direct external customers* needing training and education in the organization's product or service (i.e., client education), and *indirect external customers* seeking guidance and education on the services developed by the HRD unit (e.g., Disney seminars, Saturn programs, etc.)

Continuous Improvement

Continuous improvement is both a commitment (continuous quality improvement or CQI), and a process (continuous process improvement or CPI). The Japanese word for this second pillar is *kaizen* and is, according to Imai (1986)¹, the single most important concept in Japanese management. The commitment to quality is initiated with a statement of dedication to a shared mission and vision, and the empowerment of everyone to incrementally move toward the vision. The process of improvement is through the initiation of small, short term projects and tasks that collectively are driven by the achievement of the long term vision and mission.

The application of continuous improvement within HRD consists of two components. The first is the application within instruction through the learning and practicing of appropriate processes, tools, and skills both within the instructional setting and back on the job. The second component is the utilization of continuous improvement in the design and delivery of instruction, i.e., the continuous improvement of instruction and learning.

The process for continuous improvement, first advanced many years ago by Shewhart, and implemented by Deming is plan, do, check, and act (PDCA), a never ending cycle of improvement that should occur in both the development and delivery of instruction. It is a cyclical, iterative process. The principal educational methodology is problem-centered instruction. While no rigid rules are required for carrying out the process, the general framework of each step can be described. *Plan*, asks such key questions as, what changes are needed, what are the needed results, what obstacles need to be overcome, are data available, and what new information is needed? *Do* is the implementation of a small scale change or pilot test (i.e., instruction and training) to provide data for answers. *Check* is the assessment and measurement of the effects of the change or test. *Act* first asks does the data confirm the intended plan, are other variables influencing the plan, are the risks of going on necessary and worthwhile? Then, based on these answers, the project or task is modified and moves

Table 2. The Customers of Human Resource Development

Customers	Needs
Internal Customers - Instructional Development	
Learners (Teams and Programs/Departments)	Knowledge, skills and abilities for organizational and personal goals; joy in learning Continuous improvement, information exchange (input/output), cooperation, collaboration
Instructors/consultants	Continuous learning of the product/service, competence in instructional design and delivery
Internal Customers - Organizational Development	
Executive officers	Consultation and advisement, integration of the HRD plan within the strategic plan, participation on the Quality Council
Units/Departments, Divisions	Consultation and advisement, implementation of the HRD plan, information exchange (input/output), cooperation, collaboration.
External Customers - Direct	
Customers/clients	Training and education in using product or service
External customers - Indirect	
HRD Customers	Education on the HRD products and/or services developed by the HRD unit into the <i>Plan</i> stage again, and the iteration continues, expanding knowledge and implementing further learning. Ideally, this process would continue indefinitely - during instruction and afterwards on the job.

Speaking with Facts

A variation of this third pillar or principle is management by fact, as its application goes beyond talking about it. Speaking or managing with facts means two things. First, objective data must be collected so that the information obtained is valid. Too often "data" is based on intuition, hunches, or influence. The second thing is the implementation of decisions based on the objective data, i.e., managing according to valid information, not according to instinct or preconceptions.

As in the discussion of previous pillars speaking with facts consists of two components - the collecting of data upon which HRD strategies are developed, and the utilization of factual information in the instructional process. Within HRD speaking or managing with facts entails the collection and analysis of valid data through various means before, during and following the instructional intervention. While the need for data collection is acknowledged, the process is often short circuited with programs being implemented with minimal justifiable data. The implementation of continuous improvement is the framework within which ongoing data can be collected. Questionnaire, interviews, focus groups, validated scales, benchmarking, and environmental scanning are all appropriate means of collecting data. Effective tools for analysis of the data are also available. Lewis and Smith (1994)² identify 21 tools, including the seven basic TQ tools (fishbone diagram, checksheet, control chart, histogram, pareto diagram, run charts, and scatter diagram).

The second component is the instructional process itself, where applicable data are obtained to determine the knowledge gained, and skills learned and applied. Similar to the first component, data should be collected before, during and following the training and instruction. Before the course or seminar specific data on needed knowledge, skills and attitudes should be gathered through questionnaires and guided interviews with the targeted learners and others who will be the recipients of the training delivered. During the training data on both concepts learned, and perceptions of the learning process should be continuously gathered and analyzed, and then shared with the learners. Of critical importance is the need for flexibility within the instructional process to modify the content and delivery to accurately reflect the instruction needed and desired by the learners. Following the course or seminar ongoing data should be gathered back on the job to validate the instruction within the context of improved productivity and positive attitudes about the organization and themselves. Representative examples of such techniques will be described in the following section.

Respect for People

Respect for people within total quality management/continuous quality improvement is the acknowledgment that, while people work *at* an organization, they work *for* themselves, trying to create a meaningful and satisfying life. Recognition of this by the organization results in a framework that includes, according to Lewis and Smith (1994):

- a. Creating a sense of purpose in the workplace so that people are motivated to do their best.
- b. Keeping people informed and involved, and show them how they are a part of the bigger picture.
- c. Educating and developing people so that each individual is the best that he or she can be at what they do.
- d. Helping people communicate well so they can do their jobs with peak effectiveness.
- e. Delegating responsibility and authority downward so that people are not just "doing what I'm told," but are taking the initiative to try to make things work better.

Within the context of this paper, i.e. within human resource development, respect for people includes two groups of people - the learners (employees) and the instructional staff of the organization. While the learners are also recognized as customers, the respect, caring and support described above applies to learners. Creating an instructional environment that facilitates self-directed learning encourages respect for the learner.

Similarly, recognition and respect for the HRD staff can result in more effective learning experiences for the organization. Involvement in the strategic planning process, being asked to develop an organizational HRD plan, and being given the necessary resources to implement the plan, these all acknowledge the respect of the HRD staff, and in turn contributes to the productivity of the organization.

Application of the Four Pillars

A primary premise of this paper is, if an organization is incorporating the principles and practices of total quality management/continuous quality improvement (TQM/CQI), they must also implement within the education and training program the principles and practices of TQM/CQI. This can be done within the framework of the four principles or pillars, as described above. Thus, the question can be asked, what examples or illustrations are there, if any, of the instructional application of the four pillars?

The first instructional application was in two university courses taught by one of the authors. While the schedule was the traditional 15 week course meeting three hours each week, the content provided the opportunity to develop and apply the principles. One was an introductory course in human resource development and adult education; the other was a course on implementation of total quality management. From these "pilot" efforts a number of strategies and activities were refined and are now being implemented in various

organizations HRD program, e.g. a 20 session health care management course being conducted in area hospitals. Following are ten forms, activities and instructional strategies that reflect the four pillars.

1. A complete *Study Guide* for the course is prepared giving the objectives, session outline and schedule, and related materials and readings for each session.
2. A participant/learner background sheet is completed by each person, ideally prior to the first session or at the beginning of the first session. The data is summarized and an expanded class roster is distributed to everyone.
3. A bi-weekly feedback questionnaire is completed to determine the learners' perspective of the process, content, and pace of the course.
4. A *Quality Council* is created, consisting of a representative from each of the small groups that are formed near the beginning of the course. The Council, meeting every two weeks, provides feedback from the bi-weekly questionnaire their group members completed. From their input the course is modified to address areas (knowledge and/or skills) not adequately covered, to slow or advance the general pace of the content delivery, in general assure the needs of the learners are heard and acted on. Membership on the Council is rotated, the rotation determined by each group, in order to have multiple representation on the Council, and not overburden specific individuals with Council membership tasks throughout the course.
5. In the first or second session each person is asked to complete an *Individual Strategic Plan for Personal Continuous Quality Improvement*, a three page completion questionnaire asking each person to identify the kind of learning experience they like most/least, prepare a mission and vision statement, and relate how this course relates (or does not relate) to their mission and vision statement. These are reviewed by the instructor(s), with the course being adjusted as much as reasonably possible to accommodate the persons' expectations.
6. A variety of initial (first three sessions) small group activities are done with the small groups being formed both by the participants and by the instructor(s). The participant formed groups are done to observe both the processes (ease in forming small groups) and the product (e.g., do the same people always group together?). The instructor formed groups are structured to meet previously unknown people.
7. After approximately three sessions permanent groups are formed through a consensus process involving both the instructor and the participants, unless the participants were sent to the course as small groups. During this process a major part of one session is devoted to the meaning and purpose of teams, the difference between groups and teams, and the intent that the groups being formed will develop into teams.
8. Heavy emphasis throughout the course is placed on experiential exercises and participative activities through which the course content (knowledge) is examined and discussed, and desired skills are learned and practiced. For the most part, all exercises and activities are done within a small group format.
9. Use of case studies, both published case studies and cases drawn from the organization's experience are an integral part of the course to provide more realistic performance assessments, and assessments of employees' skills and knowledge (hopefully improved) back on the job by their supervisors. All case studies are done within a small group format.
10. Continuous feedback to each person and each team is provided throughout the course by other participants, and by the instructor(s). This is provided in the form of formal feedback, written and verbal, and informal encouraging and supportive comments. An example is the video taping of key presentations of teams, and the opportunity for feedback by the participants and instructor(s).

Conclusion

When the application of total quality, within the framework of the four pillars, is fully implemented, three outcomes occur. First, the entire course is highly experiential. The entire learning experience is designed to engage the learners in both the prepared activities, and the solicitation of the perceptions throughout the course. Adjustments are made to the course content and delivery based on the participants' collective input.

Second, within the context of the content the course is continuously focused on performance outcomes. The content determines the specificity of the behavioral performance following any learning experience. Specific performance is more readily observed for technical courses, e.g., developing computer skills, than for broader, less technical courses, e.g., health care management. Acknowledging this, the continuous focus on the implementation of the four pillars results in the extensive use of experiential learning, mentioned above, and continual feedback within the framework of the contents' application to work performance, both while the course is going on and after the course is concluded.

Third, this approach, when applied within organizations is not viewed as a separate entity, but within the systemic framework of the organization. While the content (knowledge and skills) is established, the determination of its delivery and the selection of the learners must be done within an overall plan (the HRD Plan) that complements the mission and vision of the organization. Performance technology and total quality management are co-advocates of "just-in-time" training to assure that training is both needed and will be favorably received.

A result of this entire process, this focus on respect for the customer, speaking only with facts, continuous improvement, and respect for persons, is that no two human resource development programs/activities are exactly the same. When the implementation of these four principles occur with the dynamic of an organization's mission and systemic framework, the instructional program must reflect both the immediate needs the learning experience is intended to fulfill, and contribute to the overall enhancement of the organizational environment, i.e., its culture.

Of import to the authors is this view of instructional development and learning within the context of evaluation. Presently, the principal components of evaluation are formative and summative evaluation. What is needed is to add to the framework of formative and summative evaluation the concept of "implementative" evaluation, or formative evaluation and revision during implementation that goes beyond one-to-one, small group, and field trial strategies. It is the hope of the authors that the approach to instructional development and delivery within organizations will also contribute to an expansion of the framework within which evaluation of instruction is viewed and applied.

1. Imai, Masaaki (1986). *Kaizen: The Key to Japan's Competitive Success*. New York:Random House.
2. Lewis, Ralph G, and Smith, Douglas H. (1994), *Total Quality in Higher Education*. Delray Beach, FL:St. Lucie Press, pp. 97-99.

Using 360-Degree Feedback to Evaluate Management Development: New Data, New Insights

Gary N. McLean, University of Minnesota, St. Paul

Marcia R. Sytsma, Personnel Decisions, Inc., Minneapolis

Kathryn Kerwin-Ryberg, University of Minnesota, Minneapolis

Managers in several organizations received 360-degree feedback using a standardized instrument prior to participation in a leadership development workshop, followed by individual developmental activities for up to two years, when 360-degree feedback was again provided. For the 73 individuals for whom data exist at both times, correlations among the four feedback groups (self, subordinates, peers, supervisors) were low, with the highest correlation being .40 and .29 at Times 1 and 2, both on self-subordinate comparisons. While 360-degree feedback provides multiple perspectives, its use in evaluating management development requires the appropriate context and examination of individual items against training content and objectives.

Kirkpatrick (1959a,b, 1960a,b) established four levels for the evaluation of training and development activities: Reaction, Learning, On-the-Job Behavior, and Organizational Impact. His approach has now become well-entrenched in the human resource development literature. For any one of the levels to have content validity, the operational definition of the level must be compatible with the objectives of the training. Level three, on-the-job behavior, has traditionally been measured through supervisory rating scales, similar to those used in performance appraisals, though perhaps more focused.

Supervisor ratings have limitations that challenge their validity and utility as a source for evaluating management development activities. One criticism of relying solely on the supervisor's rating is that the supervisor may have too few opportunities to observe the ratee's performance and, therefore, is not able to offer a valid rating. An alternative has been to use 360-degree feedback--the process of receiving perceptions of one's performance from a variety of sources, typically supervisor(s), peers, direct reports, self, and even customers--to measure on-the-job behavior.

Proposing an Alternative: Reasons for Using 360-Degree Feedback

While the use of 360-degree feedback is on the increase, such a "formalized program of feedback to managers from their constituents is in its infancy" (Hedge & Borman, in press). One source of motivation for the increased use of 360-degree feedback in human resource development is the increasing criticism directed toward the use of performance appraisal as a tool for development.

With special acknowledgement for their contributions to Sarah Hezlett, Keith Halperin, Joy Fisher Hazucha, Lou Quast, and Tim Hansen, all of Personnel Decisions, Inc.

(c) copyright, Gary N. McLean and Personnel Decisions, Inc., 1995

Lawler (1995), for example, summarized the situation:

The problem—and it is well-documented—is that most performance-appraisal systems do not motivate individuals nor guide their development effectively. Instead, they cause conflict between supervisors and subordinates and lead to dysfunctional behaviors. These dysfunctions often are exaggerated when performance appraisals are tied to traditional merit pay systems. They are particularly severe when the system forces supervisors to compare subordinates. (p. 29)

The argument often made for 360-degree feedback is that the administrative reasons for performance appraisals will not disappear, yet a system is needed that does not have the "rater errors, biases, and other inaccuracies" that detract from the ability of performance appraisals to guide development or the validity of the information they contain (Hedge & Borman, in press). Hedge and Borman (in press) identify three primary uses of 360-degree feedback in organizations:

- * Making general personnel decisions, such as promotions and terminations,
- * Identifying training and development needs, pinpointing employee skills and competencies that are currently inadequate but for which programs can be developed, and
- * Serving as a criterion against which selection and development programs are validated.

The second use (identifying development needs) may be one of the most frequent applications of 360-degree feedback (Kam, Howard, Antman, Cohen & Coker, 1994; Van Velsor & Leslie, 1991). This paper discusses the use of a 360-degree feedback instrument designed to facilitate management development to evaluate the effectiveness of a training event, in this case, a management development program. This use is a version of the third use identified (and listed above) by Hedge and Borman (in press). The questions are, (1) How can a 360-degree feedback instrument be used to evaluate a management development program? and (2) What are the methodological limitations of applying 360-degree feedback in this manner?

Advantages of Peer, Subordinate, and Self Ratings

Researchers have outlined advantages associated with each perspective in the 360-degree feedback process. Further, there are perceived advantages to using the multiple inputs that occur in a 360-degree feedback process as compared with the singular input from a supervisor in the more traditional evaluation systems. Latham (1986) argued that the inclusion of peers, who are more likely to have daily interactions, improves ratings. Dunnette (1993) concluded that self descriptions possess accurate components and suggested that it is premature to presume that the ratings of others should always be used to assess the validity of self-ratings. Latham and Wexley (1981) argued for the use of self-ratings from a developmental perspective. They suggested that self-ratings force the employee to focus on what's expected in the job. Also, they allow the supervisor to see how subordinates perceive their level of effectiveness.

From a review of the literature, Pollack and Pollack (1994) provided several arguments in support of subordinate ratings:

- 1) they typically have frequent contact with the manager, giving them many opportunities to observe the manager's performance, 2) they see different behaviors than supervisors and peers, giving them a unique perspective on the manager's performance, and 3) several subordinates' ratings can be averaged to produce a more reliable assessment of the manager's performance. (p. 12)

Pollack and Pollack (1994) concluded that

subordinates and peers provide a unique perspective on job performance that the supervisor of that employee may not have. The research literature also showed that subordinate and peer ratings are reliable and valid predictors of job performance. (p. 3)

Hazucha, Gebelein, Holt, and Olsen (1993), in an assessment of the psychometric properties of the PROFILOR™, a 360-degree feedback instrument developed by Personnel Decisions, Inc., concluded:

1. The more raters, from the more perspectives, the less likely the ratings are to be biased from one person's idiosyncrasies.
2. Raters should be people who have valuable observations to contribute based on their work with the person....people are likely to work most closely with their direct reports and their boss.
3. When possible, managers seeking feedback should choose raters whose perceptions are important (but not necessarily favorable) because the manager continues to rely on their cooperation.
4. We expect less-than-perfect agreement among raters because each has different goals and motivations and observes different behaviors. (pp. 29-30)

Disadvantages of Peer, Subordinate, and Self Ratings

There are, however, problems associated with the use of 360-degree feedback for training evaluation. While these problems may not be a serious issue when 360-degree feedback is used for development planning and may, in fact, help diagnose areas to emphasize during development, they pose problems when we attempt to use 360-degree feedback for other purposes (such as program evaluation or validation). One problem is in interpreting the ratings when the various groups doing the ratings provide different feedback. Which rating should predominate when they differ? Should the supervisor's rating predominate, as has been the case historically? Or should more attention be paid to the rating of those with whom the ratee has the greatest contact and, therefore, the most opportunity to observe the relevant behaviors? Or should the self-rating predominate, especially for developmental purposes, because it has the greatest potential for acceptance by the ratee?

McEvoy and Buller (1987) found that peer ratings are not well accepted by raters or ratees except when they are used for developmental purposes. On the other hand, Hedge and Borman (in press) concluded that self-ratings are "unreliable, biased, and inaccurate compared to other rating sources." Hedge and Borman (in press) also concluded that, "despite evidence that use of subordinate appraisal is on the increase, there is almost no empirical research to support upward appraisal for any purpose." They went on to argue that "the biggest potential problem with subordinate appraisals is the fear that supervisors will exact retribution if an honest, but unfavorable, appraisal is received." A recent study (Antonioni, in press) explored the "extent to which managers improved their supervisory behaviors as a result of receiving and reviewing written upward appraisal reports (feedback) with their immediate supervisors." He found that there were no differences between "managers who receive feedback with or without a performance review" and those "managers who did not receive feedback." In fact, "mean supervisory behaviors scores were lower at time two than at time one."

Other psychometric issues regarding multiple ratings include leniency tendencies, divergent and discriminant validity, and halo of self, peer, and supervisory ratings. Through a meta-analysis of the literature, Kraiger (1986) concluded, first, that, "while there is a slight leniency effect for self ratings, the effect is not so large that it would cause severe range

restriction problems in validation analyses" (p. 10). Further, since no external criteria were used in any of the studies reviewed, he concluded, second, that "any of the three sources would be equally well suited for validation...[though] all three may be good or all three may be bad" (p. 10). Finally, "peer and supervisory ratings may be more prone to halo than self ratings, but again the size of the effect is not so large that it would be a problem in validation studies, particularly when rating dimensions are combined into a single overall criterion" (p. 10). Kraiger (1986) noted, as part of the study cited earlier, that "the lack of discriminant validity in all three sources [peer, self, supervisors] revealed by the multitrait-multirater analysis would argue for the formation of composites in most instances" (pp. 10-11). "Convergent validity appears greatest between peer and supervisory ratings. Self ratings typically show a higher mean than peer or supervisory ratings which in turn are nearly equal" (p. 11).

Using 360-Degree Feedback to Evaluate Management Development

Since many 360-degree feedback instruments cover managerial skills (Van Velsor & Leslie, 1991), a natural progression to using ratings at two time periods on these instruments for program evaluation appears warranted. One such study, in which the development activities were directly related to skills and behaviors measured via 360-degree feedback, was conducted by Hazucha, Hezlett, and Schneider (1993). They investigated

changes in the skill levels of managers which occurred two years following the managers' participation in 360-degree feedback. The major objective was to investigate how skill development, development efforts, and environmental support for development are related. The Management Skills Profile (MSP) was used to measure skills and to give feedback. Of the original 198 managers who received feedback, 48 accepted the opportunity to complete the instrument again two years later. (p. 325)

Participants' skills increased during the two years, and self-other agreement was greater at Time 2 than at Time 1. Those with greater skill at Time 1 were more likely to have advanced. While providing evaluative information, the change in scores at Time 2 are not clearly as a result of either having received feedback at Time 1, having received the training provided at Time 1, nor the developmental activities that occurred between Time 1 and Time 2.

Unfortunately, problems may arise when using 360-degree feedback, or any behavioral ratings, to imply that changes occurred due to management development activities or any OD-type intervention. The first problem arises because of our dependence on consistent or convergent information from the sources we use as raters. Kraiger (1986) found correlations between perspectives on five dimensions (interpersonal skills, personal qualities, job aptitude, leadership and instruction, and level of performance) using various rating scales were greatest between peers and supervisors, but these ranged only from .32 to .52. Between self and supervisor, the correlations ranged from .16 to .30, and self-peer correlations ranged from .06 to .30. Clearly, intersperspective reliability is low. Williams and Levy's (1992) research supported some of these findings. The correlation between self- and supervisor ratings was .26, and the self-ratings showed greater leniency than did the supervisors' ratings. Hazucha, Szymanski, and Birkeland (1992) found self-supervisor correlations on 19 managerial dimensions ranging between .16 (Personal Adaptability) and .50 (Financial and Quantitative), with most of the correlations in the .2-.3 range. Finally, in another meta-analysis, Harris and Schaubroeck (1988) found peer-supervisor ratings to be correlated at .62, with self-supervisor ratings at .35 and self-peer at .36. As all of these studies indicated, intersperspective reliability ranges from .06 to .62, indicating, at best, a modest level of agreement among perspectives on a manager's performance. This study is an investigation of the degree of intersperspective reliability as a limit to evaluating a training program.

Methodology

The participants in the study reported in this presentation were individuals who had completed (or participated in) a week-long leadership training program at Personnel Decisions, Inc. during a period from 1984 to 1991. The training program was the Leadership Development Program (LDP), "an intensive 6-day training program developed by the Center for Creative Leadership in Greensboro, North Carolina" (Personnel Decisions, Inc., 1993, p. 1). The workshop is promoted as an opportunity to build leadership skills through leaderless group exercises, lecture, and feedback on test and job performance. Prior to the workshop, all participants completed the Management Skills Profile (MSP), a 122-item inventory designed to assess skills and identify development needs across 19 managerial dimensions (see Table 1). This is the same instrument used in Hazucha, Szymanski, and Birkeland (1992) and Hazucha, Hezlett, and Schneider (1993).

Table 1
Management Skills Profile Dimensions and Definitions (Hazucha, Szymanski, & Birkeland, 1992, p. 6)

Dimension	Definition
Planning	Setting goals and developing strategies and schedules for meeting those goals; anticipating obstacles and defining alternative strategies
Organizing	Scheduling and coordinating work of others; setting priorities; establishing efficient work procedures to meet objectives
Personal Organization & Time Management	Allocating one's own time efficiently; arranging information systematically and processing paperwork and other information effectively without getting bogged down in detail
Informing	Letting people know of decisions, changes, and other relevant information on a timely basis.
Oral Communication	Speaking effectively one to one and in groups; making effective presentations
Listening	Demonstrating attention to, and conveying understanding of, the comments or questions of others
Written Communication	Writing clearly and effectively; using appropriate style, grammar and tone, in informal and formal business communications
Problem Analysis & Decision Making	Identifying problems; recognizing symptoms, causes and alternative solutions, making timely, sound decisions even under conditions of risk and uncertainty
Financial & Quantitative	Drawing accurate conclusions from financial and numerical material and applying financial principles and numerical techniques to management problems
Human Relations	Developing and maintaining smooth, cooperative working relationships with peers, subordinates, and superiors; showing awareness of, and consideration for, the opinions and feelings of others.

Conflict Management	Bringing conflict or dissent into the open and using it productively to enhance the quality of decisions, arriving at constructive solutions while maintaining positive working relationships
Leadership Style & Influence	Taking charge and initiating actions, directing the activities of individuals and groups toward the accomplishment of meaningful goals and commanding the attention and respect of others
Motivating Others	Creating an environment in which subordinates and others are rewarded for accomplishment of group and individual goals
Delegating & Controlling	Clearly assigning responsibilities and tasks to others and establishing effective controls, ensuring that employees have the necessary resources and authority, and monitoring progress and exercising control.
Coaching & Developing	Evaluating employees, providing performance feedback, and facilitating professional growth
Personal Motivation	Displaying a high energy level, working long and hard to get things done, and seeking increased responsibility on the job
Personal Adaptability	Responding appropriately and competently to the demands of work challenges when confronted with changes, ambiguity, and adversity, or other pressures
Occupational & Technical Knowledge	Applying the knowledge and skills needed to do the job, including technical competence in one's own field and familiarity with policies and practices of the organization and the industry
Results Orientation	Producing high quality work, managing productively, getting results

The MSP, developed in 1984 by Personnel Decisions, Inc., is completed by the manager being assessed, and also by his or her immediate superiors, subordinates, and peers. Holt and Hazucha (1991) reported on a number of studies supporting the reliability and validity of the instrument. Longitudinal studies over a period of one to two years, with different raters, showed reliability in the .5-.6 range for non-self ratings, and from .36 to .66 for self-ratings—a result somewhat lower than those found for personality scales, as would be expected because of the types of skills and behaviors measured by the MSP. Internal consistency reliability (assessed using Cronbach's alpha) ranged from .70 to .91 across the 19 scales, with an average of .83. Finally, interrater correlations for pairs of supervisor raters ranged from .30 to .52; for subordinates, .28-.48; and for peers, from .21-.37. Average interrater reliability ranged from .28-.47. "These correlations between pairs of raters translate to an average reliability of .74 for the standard composite of 7 raters" (p. 9). Validity was originally established through content analysis. Subsequent criterion validity has been shown in a number of studies relating MSP results to performance and/or organizational level. In the current study, overall Cronbach alphas were extremely high, ranging from .96 to .99, affirming the strong internal consistency of the MSP.

Participants were primarily those who had enrolled in public offerings of the program. A small number of participants attended the program during deliveries of the training at single

organizations. LDP is a feedback-intensive training program intended to promote self-insight. It contains psychological assessment, lecture and exercises. A follow-up study was designed and carried out as an evaluation at both levels 1 and 3 (Personnel Decisions, Inc., 1993).

The LDP Follow-up Survey asked [managers] about a range of topics...their assessment of the LDP program, and to tell us about themselves, their organizations, and their coworkers in light of management development. To obtain the managers' overall reaction to the program, we asked them to give an overall assessment of the program's value and impact, as well as to assess the value of specific components of the program. (Personnel Decisions, Inc., 1993, p. 1)

Managers were also asked to provide estimates of their skill change using a scale of 1-5 (Not at All, to A Very Great Extent).

The surveys were mailed out to all participants who had attended LDP at least one year before the study began. The intention was to ensure that participants would have had a reasonable amount of time back on the job to make a considered assessment of their experiences in the program. Two organizations who had sent a large number of employees to in-house presentations of the program declined to participate in the study because they were planning similar follow-up activities. Surveys were sent out with a cover letter by PDI's Vice President of Training and Development.

Respondents who completed the survey and returned it within the specified time frame were given the opportunity to take the MSP again at no charge. Of the 317 respondents to the survey, 190 asked to retake the MSP. Of these, complete Time 1 data were available for 168 managers, and complete Time 2 data were available for 80 managers. Matched Time 1-Time 2 MSP data were available for 73 managers; the findings are based on the results of these 73 participants.

Findings and Discussion

In general, the 317 responding managers

continued to have a favorable opinion of the program. The element of the program they found most valuable was feedback....These 317 managers clearly value LDP and rank it as one of the best training programs and on par with special assignments, mentoring, and feedback as having a strong positive impact on their career. (Personnel Decisions, Inc., 1993, p. 16)

Participants said that LDP helped them be more effective (3.5), give feedback to others (3.4), be somewhat successful in implementing skills learned at LDP (3.0). Direct change estimate was 3.5. When asked how their participation in LDP influenced their career, 70.7% of respondents indicated that it had had a "positive influence," while 13.7% indicated a "strong positive influence" (p. 2). Further, from 11 possible sources of contributors to "recent career progress," the LDP experience was rated highest, an unusual situation as "most frequently managers do not rate training as a major contributor to career success" (p. 3). Self-assessment, then, was that the LDP was a success.

Several limitations in the data exist, however, restricting the statistical analyses that can be conducted. The number of participants in the final data set was low compared with the number of initial participants in the program, and the data sets for the surveys returned in the follow-up were incomplete. The number of respondents compared with the number of intervening variables

reduced the power of the statistical analyses. Perhaps most important, the MSP is a measure of behaviors, whereas the thrust of the LDP was in developing feedback seeking and application, behaviors not directly measured by the MSP.

In spite of these limitations, it was still possible to look at the data for insight into the question of using 360-degree feedback for evaluating management development activities. Interperspective reliability was examined at both times when the MSP was administered.

Correlations Across Rating Groups. Correlations by MSP dimension were run among the four rating groups at both rating times to compare intersperspective correlations. This analysis yielded 114 correlations for each time, for a total of 228 correlations. Given the number of correlations calculated and the low correlation required for statistical significance, the more critical question for analysis has to do with practical rather than statistical significance. That is, a correlation of only .30 is statistically significant at the .01 level, yet the actual relationship explained is low. A review of all correlations calculated revealed that the highest correlation obtained for Time 1 was .40, a level generally considered to be low. For Time 2, the highest correlation was even lower at .29. The highest correlations in each of the ratings group comparisons are shown in Table 2. These results are in the mid-range of intersperspective reliability found by Kraiger (1986), Hazucha, Szymanski, and Birkeland (1992), and Williams and Levy (1992). They are at the low range of those found in Harris and Schaubroeck's (1988) meta-analysis.

Table 2
Highest Correlation within Each Ratings Group Comparison at Time 1/Time 2

Group	Self	Peer	Subordinate
Peer	.24/.08		
Subordinate	.40/.29	-.23/.12	
Supervisor	.36/-.15	.28/.21	.22/.20

The correlations show clearly that there is very little relationship among the various groups in terms of how they completed the MSP. The perspective of each group is very different, and each group appears to view performance in each of the 19 dimensions in a different way from the other groups. The Time 1 self-other findings are consistent with those of Peterson (1992), though the peer-supervisor relationship is considerably lower than the .62 found by Peterson. Time 2 correlations are consistently lower. In his review of the literature, Peterson concluded that

the relationship between self-ratings and observer ratings finds that self (S) and observer (O) ratings on many individual difference variables—whether personality, ability, or work performance—show moderate agreement at best. Typically, this lack of correspondence is viewed as a problem. Another explanation, however, is that we should *expect* systematic differences (Tellegen, 1985). Simply put, S and O have access to different information regarding a target's behavior. (p. 2)

Further, since any evaluation of performance is subjective (as underscored by Scholtes (1990), changing the players doing the evaluations may lead to different ratings, even if no change occurs. The low correlations found across all dimensions within each of the groups add considerable weight to the argument that any such evaluative rating will have low reliability.

The low correlations among the rating groups underscore the difficulty of deciding whose perspective to use in evaluating the training program. This same question must be raised for other applications of 360-degree feedback, as well. For developmental purposes, it is helpful to be able

to view one's own performance from the perspective of several groups. But when a decision needs to be made, such as evaluating a developmental activity or making an administrative decision (pay, promotion), varied and contradictory evidence creates confusion rather than options. Guidelines are needed to determine the most valid source for a particular dimension, behavior, or application.

Earlier in this paper, we mentioned problems with using 360-degree feedback to evaluate training programs due to the potential for low reliability and convergence between rating sources, a finding replicated in this study. Another limitation of the typical 360-degree feedback instrument and other direct ratings of individual performance arises from the methods needed to translate these "snapshots" of performance or static ratings into measures of a dynamic entity such as change.

These observations are similar to those outlined by Peterson (1993) in his presentation of the results of a longitudinal study investigating behavior change in an individualized coaching program. Peterson found that people with different perspectives (such as the boss, participant, and coach) varied in the amount of change they observed on a measure similar to 360-degree feedback instruments but showed much greater agreement when they were asked to rate directly the amount of change they observed in the participants (Peterson, 1993). Peterson explained these results in the context of alpha, beta, and gamma change provided by Golembiewski, Billingsley, and Yeager (1975).

Golembiewski et al. provided the following definitions: "Alpha Change involves a variation in the level of some existential state, given a constantly calibrated measuring instrument related to a constant conceptual domain" (p. 134).

Alpha change is the type of change we hope for in a developmental activity, where the behavior is directly affected by an intervention. In contrast, "Beta Changes involve the recalibration of the intervals used to measure some stable dimension of psychological space, as in preintervention vs. post-intervention responses" (Golembiewski et al., p. 135), and "Gamma Change...involves the basic redefinition of the relevant psychological space as a consequence of an OD intervention" (p. 137).

Beta change occurs when the intervention increases or decreases rating biases affecting the scale intervals, such as leniency, stringency, etc. Gamma change is reflected by a new understanding of the norms against which ratings are made, or by learning about the expectations of others. It is quite possible for beta and gamma changes to keep us from observing alpha changes when we simply compare ratings given at two points in time. Peterson's "retrospective degree of change rating" (Peterson, 1993, p. 2) is one tool to get around these potentially conflicting types of change.

Conclusions and Recommendations

While the desire to obtain level-3 evaluation of a training intervention in Kirkpatrick's model is important and valid, the use of 360-degree feedback, in its current state of development, has limitations for evaluating the training program described here. The use of 360-degree feedback seems well established when used for individual developmental purposes. Trying to use it for evaluative or administrative purposes reflects the frustration of those who continue to work at improving performance appraisal systems at an organizational level instead of recognizing the inherent deficiencies in the system itself. The challenge to human resource development, then, is to determine explicitly the outcomes that are expected from management development activities and design a program evaluation system that responds specifically to these outcomes. Further, given the context-specific nature of performance feedback, it may be essential to view evaluation on an organization-specific basis, thus avoiding commingling of organizations in evaluation research. Another future direction for research is to establish within organizations the most valid

sources for various skill or performance dimensions. In summary, 360-degree feedback needs to be applied very cautiously as a tool to evaluate training programs.

References

- Antonioni, D. (in press). Problems associated with implementation of an effective upward appraisal feedback process: An experimental field study. Human Resource Development Quarterly.
- Dunnette, M.D. (1993). My hammer or your hammer? Human Resource Management, 32, 373-384.
- Golembiewski, R.T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. Journal of Applied Behavioral Science, 12, 133-157.
- Harris, M.M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. Personnel Psychology, 41, 43-61.
- Hazucha, J.F., Gebelein, S.J., Holt, K.E., & Olsen, H.E. (1993, May 1). Using 360-degree feedback to enhance management development. Paper presented at the Eighth Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco.
- Hazucha, J.F., Hezlett, S.A., & Schneider, R.J. (1993). The impact of 360-degree feedback on management skills development. Human Resource Management, 32(2-3), 325-351.
- Hazucha, J.F., Szymanski, C., & Birkeland, S. (1992, August). Will my boss see my ratings? Effect of confidentiality on self-boss rating congruence. Paper presented at the Annual Meeting of the American Psychological Association, Washington, DC.
- Hedge, J.W., & Borman, W.C. (in press). Changing conceptions and practices in performance appraisal. In A. Howard (Ed.). The changing nature of work. San Francisco: Jossey-Bass.
- Holt, K.E., & Hazucha, J.F. (1991). Management skills profile technical summary. Minneapolis: Personnel Decisions, Inc.
- Kam, S.M., Howard, S., Antman, D., Cohen, L., & Coker, K. (1994, January). 360-degree feedback: A benchmarking study. Kaiser Permanente Technical Report. Oakland, CA: Kaiser Permanente Consulting, Research and Development.
- Kirkpatrick, D. (1959a). Techniques for evaluating training programs. Journal of the American Society of Training Directors, 13(11), 3-9.
- Kirkpatrick, D. (1959b). Techniques for evaluating training programs. Journal of the American Society of Training Directors, 13(12), 21-26.
- Kirkpatrick, D. (1960a). Techniques for evaluating training programs. Journal of the American Society of Training Directors, 14(1), 13-18.
- Kirkpatrick, D. (1960b). Techniques for evaluating training programs. Journal of the American Society of Training Directors, 14(2), 28-32.
- Kraiger, K. (1986, April 10). An analysis of relationships among self, peer, and supervisory ratings of performance. Paper presented at the First Annual Conference of the Society of Industrial/Organizational Psychologists, Chicago, IL.
- Latham, G.L. (1986). Job performance and appraisal. In C.L. Cooper & I. Robertson (Eds.). International review of industrial and organizational psychology (Vol. 1). Chichester, England: John Wiley.
- Latham, G.L., & Wexley, K.N. (1981). Increasing productivity through performance appraisal. Reading, MA: Addison-Wesley.
- Lawler, E.E., III. (1995). Performance management: The next generation. Quality Digest, 15(2), 29-31.
- McEvoy, G.M., & Buller, P.F. (1987). User acceptance of peer appraisals in an industrial setting. Personnel Psychology, 40, 785-797.

- Personnel Decisions, Inc. (1993). Leadership development program. Follow-up survey. Minneapolis: Author.
- Peterson, D.B. (1992, August). Personality predictors of correspondence between self-report and observer ratings. Paper presented at the Annual Conference of the American Psychological Association, Washington, DC.
- Peterson, D.B. (1993, April). Measuring change: A psychometric approach to evaluating individual training outcomes. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.
- Pollack, L.J., & Pollack, D.M. (1994, June 26-30). Using 360-degree feedback in performance appraisal. Paper presented at the International Personnel Management Association Assessment Council Conference, Charleston, SC.
- Scholtes, P.R. (1990). An elaboration on Deming's teachings on performance appraisal. In McLean, G.N., Damme, S.R., & Swanson, R.A. Performance appraisal: Perspectives on a quality management approach (pp. 24-52). Alexandria, VA: American Society for Training and Development.
- Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In A.H. Tuma & J.D. Maser (Eds.), Anxiety and the anxiety disorders. Hillsdale, NJ: Erlbaum. (Cited in: Peterson, 1992.)
- Van Velsor, E., & Leslie, J.G. (1991). Feedback to managers: A review and comparison of sixteen multirater feedback instruments. Greensboro, NC: Center for Creative Leadership.
- Williams, J.R., & Levy, P.E. (1992). The effects of perceived system knowledge on the agreement between self-ratings and supervisor ratings. Personnel Psychology, 42, 835-847.